University of Peradeniya

Cluster Analysis for Metagenomics Binning

Delini S.[1]
Damayanthi Herath[1]
[1] *Department of Computer Engineering, Faculty of Engineering, University of Peradeniya.*

Abstract- Binning, a critical step in metagenomic analysis, involves grouping nucleotide sequences from similar species. This study explores data-driven approaches to enhance binning efficiency, particularly focusing on a dissimilarity-based method. The proposed approach increases the number of contigs binned compared to conventional methods while maintaining acceptable accuracy levels.

Background- A key bioinformatic step in metagenomics is binning which refers to grouping nucleotide squences belonging to similar species. This work considers metagenomic experiments involving shotgun sequencing. It identifies the key limitations with the two-tiered binning approach (2T binning method) and aimed to develop a refining strategy to improve the binning performance.

Methodology

- Features used in two-tiered binning methods are used by the proposed method.

- **OFDEG**
- **GC Content**
- **Tetranucleotide frequency**

- Noise handling techniques are also used in this method.

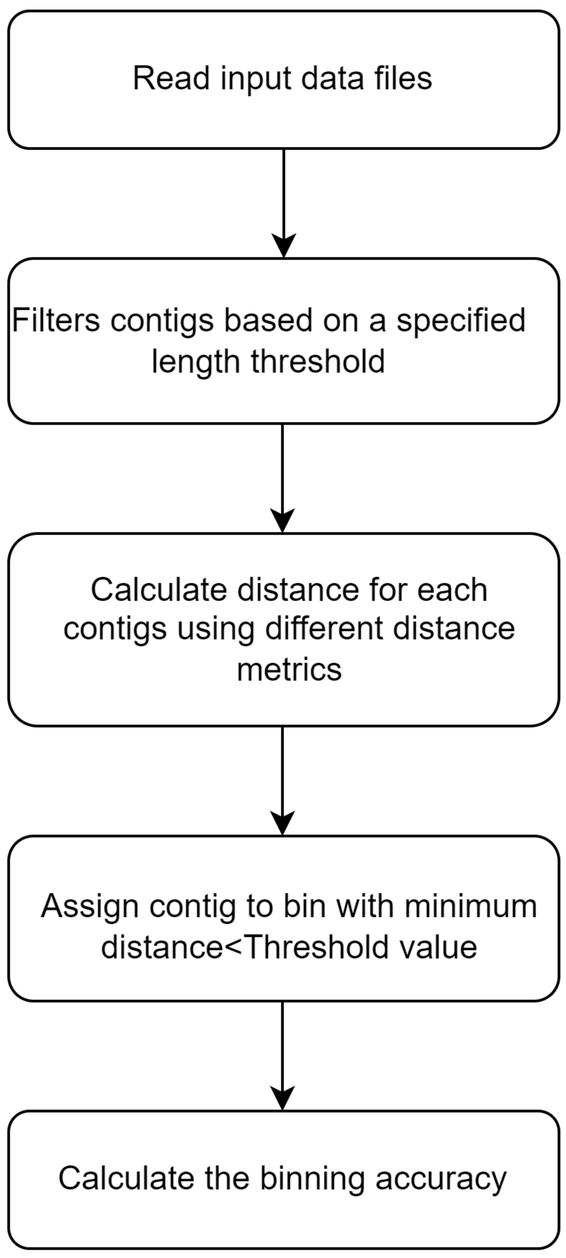

Figure 1: workflow of the proposed method

A new parameter was created to indicate the likelihood of contig belonging to a particular bin. The parameter is generated using Mahalanobis, Chebyshev, Manhattan and Euclidean.

Mahalanobis -Distance between two vectors is calculated by taking the square root of the transpose of the difference vector multiplied by the inverse of the covariance matrix, and then the difference vector itself. $D^2 = (x-m)^T \cdot C^{-1} \cdot (x-m)$

Chebyshev -Maximum absolute difference between corresponding elements of two vectors.

$$Chebyshev = max\left(|x_i - y_i|\right)$$

Manhattan- Sum of the absolute difference between corresponding elements of two vectors.

$$Manhattan = \sum_{i=1}^{n} |x_i - y_i|$$

Euclidean- square root of the sum of the squared differences between corresponding elements of two vectors.

$$Euclidean = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Results

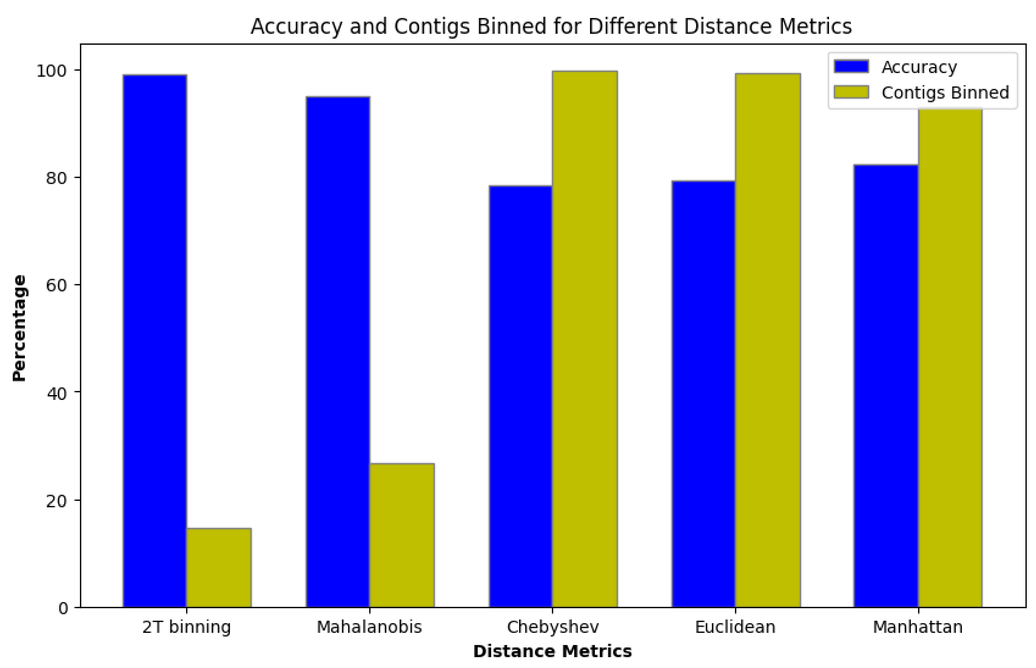

Figure 2: Percentage of Accuracy and Contigs Binned vs Different Distance Metrics

Percentage of contigs binned increased with an acceptable amount of accuracy.

Conclusion

- This study focuses on existing 2T binning method, suggesting potential enhancements.
- Using the Manhattan distance metric, the analysis achieved an accuracy of 82.35%. Moreover, 92.97% of the contigs were successfully binned.

Practical Use

Proposed binning refining method is anticipated to offer more accurate insights into microbial community composition benefiting fields such as environmental science, biotechnology, and medicine.

References

[1]https://medium.com/geekculture/7-important-distance-metrics-every-data-scientist-should-know-11e1b0b2ebe3
[2]https://link.springer.com/content/pdf/10.1140/epjc/s10052-023-12314-z.pdf

Contact details
Name : Dr. Damayanthi Herath
Tel. No.: +94779667468
Email : damayanthiherath@eng.pdn.ac.lk

Multidisciplinary AI Research Centre (MARC)
University Research Council
University of Peradeniya
Peradeniya, 20400, Sri Lanka